DOCUMENT RESUME

ED 462 418 TM 033 682

AUTHOR Ban, Jae-Chun; Hanson, Bradley A.; Yi, Qing; Harris, Deborah J.

TITLE Data Sparseness and Online Pretest Item Calibration/Scaling Methods in CAT. ACT Research Report Series.

INSTITUTION American Coll. Testing Program, Iowa City, IA.
REPORT NO ACT-RR-2002-1
PUB DATE 2002-01-00
NOTE 23p.
AVAILABLE FROM ACT Research Report Series, P.O. Box 168, Iowa City, IA 52243-0168. Tel: 319-337-1028; Web site: http://www.act.org.
PUB TYPE Reports - Evaluative (142)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Adaptive Testing; *Computer Assisted Testing; Data Analysis; Error of Measurement; Estimation (Mathematics); Maximum Likelihood Statistics; *Online Systems; *Pretests Posttests; *Scaling; Simulation; Test Items
IDENTIFIERS *Calibration; EM Algorithm

ABSTRACT

The purpose of this study was to compare and evaluate three online pretest item calibration/scaling methods in terms of item parameter recovery when the item responses to the pretest items in the pool would be sparse. The three methods considered were the marginal maximum likelihood estimate with one EM cycle (OEM) method, the marginal maximum likelihood estimate with multiple EM cycles (MEM) method, and Stocking's Method B. The three methods were evaluated using simulations of data from computerized adaptive tests (CAT). The MEM method produced the smallest average total error in recovering the 240 pretest item characteristic curves. Stocking's Method B yielded the second smallest average total error in parameter estimation. In terms of scale maintenance, the MEM method and Stocking's Method B performed well in keeping with the scale of the pretest items on the same scale as that of the true parameters. With the OEM method, the scale of the pretest item parameter estimates deviated from that of the true parameters. (Contains 1 figure, 4 tables, and 14 references.) (Author/SLD)

Reproductions supplied by EDRS are the best that can be made
from the original document.

ACT Research Report Series 2002-1

ED 462 418

Data Sparseness and Online Pretest Item Calibration/Scaling Methods in CAT

Jae-Chun Ban

Bradley A. Hanson

Qing Yi

Deborah J. Harris

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

P. Farrant

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

TM033682

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

ACT

January 2002

For additional copies write:
ACT Research Report Series
P.O. Box 168
Iowa City, Iowa 52243-0168

© 2002 by ACT, Inc. All rights reserved.

Data Sparseness and Online Pretest Item Calibration/Scaling Methods in CAT

Jae-Chun Ban
Bradley A. Hanson
Qing Yi
Deborah J. Harris

Abstract

The purpose of this study was to compare and evaluate three online pretest item calibration/scaling methods in terms of item parameter recovery when the item responses to the pretest items in the pool would be sparse. The three methods considered were the marginal maximum likelihood estimate with one EM cycle (OEM) method, the marginal maximum likelihood estimate with multiple EM cycles (MEM) method, and Stocking's Method B. The three methods were evaluated using simulations of data from computerized adaptive tests (CAT).

The MEM method produced the smallest average total error in recovering the 240 pretest item characteristic curves. Stocking's Method B yielded the second smallest average total error in parameter estimation. The OEM method yielded a large average total error in parameter estimation. In terms of scale maintenance, the MEM method and Stocking's Method B performed well in keeping the scale of the pretest items on the same scale as that of the true parameters. With the OEM method, the scale of the pretest item parameter estimates deviated from that of the true parameters.

Data Sparseness and Online Pretest Item Calibration/Scaling Methods in CAT

Calibrating pretest items is a necessary part of a large computerized adaptive testing (CAT) program. Collecting online pretest item responses and calibrating the pretest items based on those responses are a way of replenishing an item pool. Online pretest item calibration refers to estimating the parameters of the pretest items that are presented to examinees along with operational items under CAT or computer based testing (CBT). Several studies have proposed online pretest item calibration methods (Ban, Hanson, Wang, Yi, & Harris, 2000; Stocking, 1988; Wainer & Mislevy, 1990) using parametric item response functions in which pretest item characteristic curves are estimated through a specific mathematical model such as a three-parameter logistic (3-PL) item response theory (IRT) model.

When pretest items are administered with operational items, item responses on the pretest items as well as the operational items are, typically, sparse. That is, not all examinees respond to all items. The sparse item response data is one of the challenges to accurately calibrating pretest item parameters (Ban et al., 2000; Haynie & Way, 1995; Hsu, Thompson, & Chen, 1998; Stocking, 1988). Hsu et al. (1998) compared the precision of item parameter estimates obtained under the sparse matrix to that of estimates calibrated under the item-examinee full matrix. They showed that using a pretest item calibration method, the sparse data lowers the precision of item parameter estimates.

In realistic settings, sparse data may exist for both the pretest item responses and the operational item responses. When there is a pretest item pool, each examinee may take only a subset of the pretest items in the pool. If the examinees were administered different sets of pretest items, the item responses to the pretest items in the pool would be sparse. Due to the nature of a CAT, operational item responses are also sparse.

Ban et al. (2000) compared and evaluated five pretest item calibration methods with different sample sizes in terms of item parameter recovery when there was sparse data only on operational items: the marginal maximum likelihood estimate with one EM cycle (OEM) method (Wainer & Mislevy, 1990), the marginal maximum likelihood estimate with multiple EM cycles (MEM) method (Ban et al., 2000), BILOG with Strong Priors (Ban et al., 2000), Stocking's Method A (Stocking, 1988), and Stocking's Method B (Stocking, 1988). They reported that the MEM method and Stocking's Method B worked very well, and the OEM method might perform well in some situations.

This study is an extension of the Ban et al. (2000) study, which evaluates how the three online calibration methods (the OEM method, the MEM method, and Stocking's Method B) behave with sparse item response data on both pretest items and operational items. The purpose of this study was to continue to compare and evaluate three online pretest item calibration/scaling methods for sparse pretest item response data in terms of item parameter recovery.

A Brief Review of Pretest Item Calibration Methods

The OEM Method

Wainer and Mislevy (1990, pp. 90-91) described the marginal maximum likelihood estimate with one EM cycle approach for calibrating online pretest items. The OEM method takes just one E-step using the posterior distribution of ability, which is estimated based on item responses only from the operational CAT items, and just one M-step to estimate the pretest item parameters, involving response data from only the pretest items (Wainer & Mislevy, 1990). With this approach, the item parameter estimates of the pretest items would only be updated once because only one M-Step of the EM cycle is computed. One advantage of this method is that no

pretest item can contaminate other pretest items because this method calibrates the pretest items through only one E-M cycle. The pretest item parameters are in theory on the same scale as the operational item parameters because the single E-step uses a posterior distribution of ability based only on operational items.

The MEM Method

As a variation of the OEM method, Ban et al. (2000) increased the number of EM cycles until a convergence criterion was met. The MEM method is similar mathematically to the OEM method. The first EM cycle with the MEM method is the same as the OEM method. The MEM method computes the posterior distribution using the operational item responses and estimates the pretest item parameters through the first M-step. However, from the second E-step, the MEM method uses item responses on both the operational items and pretest items to get the posterior distribution. For each M-step iteration, the item parameter estimates for the operational items are fixed, whereas parameter estimates for the pretest items are updated until the pretest item parameter estimates converge. One reason for fixing the operational item parameters during the EM cycles is to prevent the possible contamination of the operational item parameter estimates by the pretest items. Through the EM cycles, however, the pretest items may affect each other's parameter estimation. With this method, the pretest items are on the same scale as the operational items because the operational item parameters are fixed in the M-steps.

In implementing the MEM and OEM methods in practice, a Bayesian modal estimation approach may be used by multiplying the marginal maximum likelihood equations by a prior distribution for the pretest item parameters in order to prevent the extreme values of item parameter estimates.

Stocking's Method B

Stocking's Method B (1988) involves two steps: the first step is to calibrate pretest item parameters and the second step is to rescale the parameter estimates using anchor items. Each examinee is administered some operational items, pretest items, and anchor items. The alternative design is that each simulee takes either pretest items or anchor items, which requires more simulees (Stocking, p. 21, 1988). Stocking's Method B estimates examinees' ability using the operational item responses. The estimated abilities are, then, fixed in order to calibrate both the pretest items and anchor items. The two sets of item parameter estimates for the anchor items, the original item parameters and the re-estimated parameters, are used to compute a scale transformation to minimize the difference between the two test characteristic curves (Stocking & Lord, 1983). This scale transformation is then used to place the parameter estimates for the pretest items onto the same scale as the operational item pool. The pretest item parameter estimates are in theory on the same scale as the operational item parameter estimates due to the scale transformation.

Method

Instrument and Data

This study used sixteen 60-item paper-and-pencil ACT Mathematics test forms (ACT, 1997) consisting of six content categories: Pre-Algebra (PA), Elementary Algebra (EA), Intermediate Algebra (IA), Coordinate Algebra (CA), Plane Geometry (PG), and Trigonometry (TG). The computer program BILOG (Mislevy & Bock, 1990) was used to estimate item parameters for all items assuming a three-parameter logistic IRT model. The estimated item parameters were treated as true parameters for CAT simulations.

Of the 960 available items, 940 items were allocated as follows: 240 pretest items (pretest item pool), 100 anchor items (anchor item pool), and 600 CAT items (operational item pool). (The remaining 20 items were not used in this study.) Descriptive statistics for the true item parameters of the operational items, pretest items, and anchor items are provided in Table 1.

TABLE 1

Descriptive Statistics for True Item Parameters of the Operational Items, Pretest Items, and Anchor Items

Item Pool Name	# of Items	*a*		*b*		*c*	
		Mean	SD	Mean	SD	Mean	SD
Operational Items	600	1.0445	0.3465	0.0844	1.0842	0.1881	0.0827
Pretest Items	240	1.0566	0.3655	-0.0051	1.1671	0.1810	0.0758
Anchor Set 1	10	1.0584	0.3287	0.0097	1.3457	0.1554	0.0904
Anchor Set 2	10	0.9587	0.2466	-0.0026	1.2307	0.2253	0.1000
Anchor Set 3	10	1.1130	0.4187	-0.0262	1.4753	0.1605	0.0774
Anchor Set 4	10	0.9551	0.3277	0.1654	0.9284	0.2132	0.1219
Anchor Set 5	10	1.0581	0.3942	0.0867	0.6662	0.2331	0.0819
Anchor Set 6	10	1.0677	0.3409	0.2278	0.7010	0.2134	0.1308
Anchor Set 7	10	0.9808	0.1962	0.1246	0.5671	0.1867	0.0612
Anchor Set 8	10	1.1126	0.3696	0.0634	0.6294	0.1323	0.0538
Anchor Set 9	10	1.2101	0.4953	0.2178	1.2401	0.1616	0.0481
Anchor Set 10	10	1.1439	0.3894	0.1247	1.7633	0.1727	0.0555

The 240 pretest items were divided into 24 pretest item sets, each set consisting of 10 pretest items. The pretest item sets were constructed to include items from six content categories (PA, EA, IA, CG, PG, and TG), if possible. The 100 anchor items were also divided into 10 sets of anchor items. Each set of anchor items were constructed to be, as close as possible, representative of the operational item pool in terms of item difficulty and content. It was nearly impossible for all anchor sets, each only consisting of 10 items, to be exactly representative in terms of difficulty and content simultaneously. Table 2 shows percentages of items in each content category for the operation item pool and each anchor item set. The anchor item sets contain 2 (20%) or 3 (30%) items in the PA and PG content categories alternatively. It can be seen that there was some discrepancy in the percentages of items in the content categories for

each anchor item set, compared to the operational item pool, due to the anchor item sets each having only ten items.

TABLE 2

Percentages of Items in Each Content Category

Item Pool or Item Set	# of Items	Content Categories					
		PA	EA	IA	CG	PG	TG
Operational Items	600	23.5	16.7	14.8	15	23.5	6.7
Each Anchor Item Set	10	20(30)	20	10	10	30(20)	10

CAT Simulation Procedures

Since true item parameters are never known in the real world, this study used true item parameters only for generating item responses and for evaluating the performance of the item calibration methods. The item parameters for the 600 operational items were estimated from a full item-simulee response matrix generated using the true parameters and 3,000 randomly selected simulees from a standard normal distribution. We call these estimated item parameters "baseline" parameter estimates. We used the baseline parameter estimates for item selection and ability estimation in the CAT simulations.

We simulated the operational item responses to a 30-item-fixed-length adaptive tests for 12,000 randomly selected simulees from the standard normal ability distribution. We derived the ability estimates using expected a posteriori Bayesian estimation (EAP; Bock & Mislevy, 1982) during the CATs. The initial prior distribution for EAP ability estimates was assumed to be normal with a mean of 0.0 and a standard deviation of 1.0. At the end of the 30-item-fixed-length tests, ability estimates were computed using maximum likelihood estimation (MLE) procedures. The simulated CAT began with an initial estimate of ability of 0.0. A random number from a uniform distribution U(0,1) was drawn to decide which content category to select.

Since the proportions of operational items in the six content categories were known, the random number was used to select content categories from which to administer items so the proportion of items administered from each content category was approximately equal to the target values. With this procedure, items to be administered were selected in a balanced way so that as nearly as possible, each simulee takes items proportional to the proportions of operational items in the six content categories. Table 3 shows observed percentages of operational items administered in each content category for one replication (with 12,000 simulees). It can be seen from comparing Tables 2 and 3 that the observed percentages of items in the six content categories are close to those in operational item pool. Similar results were observed for other replications.

TABLE 3

Observed Percentages of Operational Items Administered in Each Content Category for Replication 1

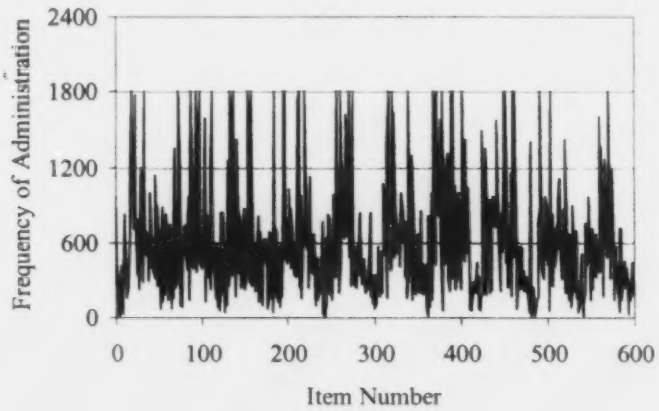
Item Pool	Content Categories					
	PA	EA	IA	CG	PG	TG
Observed Percentages	23.3	17	14.3	14.6	24.1	6.8

The matched difficulty method (Urry, 1970) was modified and used as the item selection criterion. The matched difficulty method typically selects an item that has a difficulty closest to the current ability estimate. However, in this study, once a content category was selected, absolute differences were computed between the provisional ability estimate for the simulee and the *b* parameters for all unadministered items to the simulee within the content category. Two items having the smallest and second smallest absolute differences were candidates for selection. Again, a random number from a uniform distribution U(0,1) was drawn. If the random number was less than or equal to 0.5, then the item having the smallest absolute difference was administered; otherwise, the item having the second smallest absolute difference was

administered. This randomization of two items was to protect against overexposure of items having the minimum absolute difference.

For item exposure control, we set the upper limit of the item exposure rate to 0.15, which resulted in the maximum item exposure of 1,800 for some items. The modified matched difficulty method with the upper limit of item exposure performed well in terms of item usage rate. Figure 1 shows that frequencies of administration for the 600 operational items for one replication of 12,000 simulated examinees. No item exceeded the upper limit (1,800) of item exposures. Also, there were no unused items, although positive frequencies may not discernible for some items in the figure. Similar results were observed for other replications.

FIGURE 1. Frequency of item exposure for Replication 1



The same simulees who took a fixed-length CAT were simulated to take 10 pretest items in the following way: the first 250 simulees took the pretest items 1 to 10, the next 250 simulees took the pretest items 6 to 15, and so forth. The last 250 of the 12,000 simulees were simulated to take the pretest items 1 to 5 and 235 to 240. Therefore, each block of 250 simulees took 10 pretest items in common, while two consecutive examinee blocks had 5 pretest items in common. For example, the first 250 simulees (Block 1) who took pretest items 1 to 10 and the

next 250 simulees (Block 2) who took pretest items 5 to 15 had common pretest items 5 to 10. There was a total of 48 blocks (= 12,000 / 250). In effect, the simulees took 10 of the total pretest items, and each pretest item was administered to 500 simulees (there were 500 responses to each item). This way of administering pretest items resulted in sparseness in the pretest item data matrix.

The 100 anchor items were administered to the same simulees in the following way: the first 250 simulees took anchor item Set 1, the next 250 simulees took anchor item Set 2, and so on. Since each set of anchor items was constructed to be representative of the operational item pool in terms of item difficulty and content, it was designed for each simulee to take the anchor items as a set. Unlike pretest item administration, different blocks of examinees did not have any common anchor items. The data matrix of item responses on the anchor items was also sparse.

Pretest Item Calibration Procedures

The computer simulations were performed using programs written in Visual Basic and C++. An open-source C++ toolkit for IRT parameter estimation (Hanson, 2000a) was used to implement the item parameter estimation. The OEM and MEM methods were implemented with the same procedures as the EPDIRM computer program (Hanson, 2000b) uses. Section 2.4 of the EPDIRM manual describes how to use the program to estimate pretest items using the MEM method.

When implementing Stocking's Method B, both pretest and anchor item parameters were first calibrated such that the simulees' abilities were estimated using the item responses on the adaptively administered operational items and the ability estimates were treated as true in order to estimate item parameters. A Stocking-Lord scale transformation function (Stocking & Lord, 1983) was computed using the two sets of item parameter estimates for the 100 anchor items, the

original item parameters and the re-estimated parameters. The pretest item parameter estimates were then transformed using the Stocking-Lord scale transformation function, which resulted in the Stocking's Method B item parameter estimates.

The following priors were used for both pretest and anchor item parameters: $a \sim$ lognormal (0, 0.5), $b \sim$ Beta4 (1.01, 1.01, -6, 6), and $c \sim$ Beta (5, 17). The same priors were used for the three pretest item calibration methods. Note that the prior distribution for the *b*-parameter is a four parameter beta that ranged from -6 to 6 almost uniformly. The prior distributions for the item parameters were used to prevent extreme values of the item parameter estimates.

Analyses

The simulations were replicated 100 times. The three methods were used to estimate pretest item parameters for each of the 100 data sets. This produced 100 item parameter estimates for each of the 240 pretest items for each method.

The first analysis evaluated how well the three methods maintained the scale of the true parameters. In theory, the three methods should put the estimated parameters of pretest items on the same scale as operational items. It is important to empirically examine how well the methods maintained the scale. Since the true pretest item parameters were on the same scale as the true operational item parameters, a Stocking-Lord scale transformation function (Stocking & Lord, 1983) using a standard normal distribution of ability as a weight function was computed for each replication between the estimated pretest item parameters for each method and true pretest item parameters. The average slope and intercept of the scale transformation functions are computed for each method. The information will show how well the methods produce estimates in the scale of the true item parameters.

The second analysis examined the extent to which the true item characteristic curves of the pretest items were recovered. Let $P(\theta \mid a_k, b_k, c_k)$ be the true item characteristic curve for the 3-PL logistic item response model, where a_k, b_k, and c_k are the true item parameters for pretest item k. Let $P(\theta \mid \hat{a}_{kr}, \hat{b}_{kr}, \hat{c}_{kr})$ be the estimated item characteristic curve for item k on replication r, where $\hat{a}_{kr}, \hat{b}_{kr}, \hat{c}_{kr}$ are estimated pretest item parameters. The weighted mean squared difference between the true item characteristic curve and the estimated item characteristic curve, which is called the weighted mean squared error (WMSE), for pretest item k is

$$\frac{1}{100}\sum_{r=1}^{100}\int_{6}[P(\theta \mid a_k, b_k, c_k) - P(\theta \mid \hat{a}_{kr}, \hat{b}_{kr}, \hat{c}_{kr})]^2 w(\theta)d\theta, \qquad (1)$$

where $w(\theta)$ is a weight function based on a $N(0, 1)$ distribution. The integral is approximated using evenly spaced discrete θ points on the finite interval (-6, 6) at increments of 0.1. Each finite θ point was weighted based on a normal distribution.

WMSE may be decomposed into the weighted squared bias (WSBias) and the weighted variance (WVariance):

$$\frac{1}{100}\sum_{r=1}^{100}\int_{6}[P(\theta \mid a_k, b_k, c_k) - P(\theta \mid \hat{a}_{kr}, \hat{b}_{kr}, \hat{c}_{kr})]^2 w(\theta)d\theta =$$

$$\int_{6}[P(\theta \mid a_k, b_k, c_k) - m_k(\theta)]^2 w(\theta)d\theta + \frac{1}{100}\sum_{r=1}^{100}\int_{6}[P(\theta \mid \hat{a}_{kr}, \hat{b}_{kr}, \hat{c}_{kr}) - m_k(\theta)]^2 w(\theta)d\theta, \qquad (2)$$

where

$$m_k(\theta) = \frac{1}{100}\sum_{r=1}^{100} P(\theta \mid \hat{a}_{kr}, \hat{b}_{kr}, \hat{c}_{kr}),$$

and WSBias and WVariance are the first and second terms on the right side of Equation 2. Means and standard deviations of the WVariance, the WSBias, and the WMSE across the 240 pretest items are examined.

The third analysis evaluated how well the *a*, *b*, and *c* parameters were separately recovered through Bias, standard error (SE), and root mean square error (RMSE).

$$Bias(\hat{\beta}_k) = \frac{1}{100}\sum_{r=1}^{100}(\hat{\beta}_{kr} - \beta_k), \tag{3}$$

$$SE(\hat{\beta}_k) = \sqrt{\frac{1}{100}\sum_{r=1}^{100}(\hat{\beta}_{kr} - \frac{\sum_{r=1}^{100}\hat{\beta}_{kr}}{100})^2}, \tag{4}$$

$$RMSE(\hat{\beta}_k) = \sqrt{\frac{1}{100}\sum_{r=1}^{100}(\hat{\beta}_{kr} - \beta_k)^2}, \tag{5}$$

where β_k is a true parameter *a*, *b*, or *c* of pretest item *k*, $\hat{\beta}_{kr}$ is an estimated parameter *a*, *b*, or *c* of pretest item *k* on replication *r*. Here, means and standard deviations of the Bias, the SE, and the RMSE across the 240 pretest items are examined.

Results

Scale Maintenance

Table 4 provides the average slopes and intercepts of the scale transformation functions between the estimated and true pretest item parameters for different methods, where the average is taken over the 100 replications. If the scale of estimated parameters were on the same scale as the true parameters, the intercept and slope should be 0 and 1, respectively. In Table 4, the average intercepts and slopes were close to 0 and 1 for the MEM method and Stocking's Method B. The slope of the OEM method was farthest from 1. In effect, Table 4 shows that the parameter estimates based on the MEM method and Stocking's Method B appear to be on the

same scale as the true parameters. The OEM method appears somewhat different from the scale of the true parameters.

TABLE 4

Average Scale Transformation Function Using Stocking-Lord Method

Estimation Method	Intercept	Slope
OEM	0.0080 (0.0078)	0.9038 (0.0078)
MEM	0.0048 (0.0081)	0.9938 (0.0088)
Stocking's Method B	0.0246 (0.0094)	0.9952 (0.0117)

( ) Standard Deviation over Replications

Average WSBias, Average WVariance, and Average WMSE

For each pretest item, WSBias, WVariance, and WMSE were computed. Table 5 presents average WSBias, average WVariance, and average WMSE, where the average is taken over the 240 pretest items. The MEM method produced the smallest average WSBias. Stocking's Method B produced the second smallest average WSBias. The OEM method produced the largest average WSBias in this study. It appears that the MEM method outperformed the other methods in terms of the systematic error value (i.e., WSBias). These results were consistent with the results in Ban et al. (2000).

In terms of average WVariance, the OEM method yielded the smallest value, Stocking's Method B produced the next smallest, and the MEM method produced the largest average WVariance. However, the difference in the average WVariance between Stocking's Method B and the MEM method was very small.

The average WMSE is a sum of the average WSBias and average WVariance. The MEM method produced the smallest average WMSE, Stocking's Method B followed next, and the

OEM method produced the largest average WMSE, which was consistent with the results in Ban et al. (2000).

In sum, with the item characteristic curve difference criterion, the MEM method appeared to perform best, Stocking Method B was second best, and the OEM method performed worst in this study.

TABLE 5

Average WSBias, Average WVariance, and Average WMSE

Estimation Method	WSBias	WVariance	WMSE
OEM	0.4287 (0.3290)	0.6720 (0.1431)	1.1007 (0.3814)
MEM	0.0926 (0.1111)	0.7670 (0.1718)	0.8596 (0.2250)
Stocking's Method B	0.1562 (0.1374)	0.7523 (0.1702)	0.9085 (0.2462)

( ) Standard Deviation over Pretest Items

Average Bias, Average Standard Error, and Average Root Mean Square Error of Item parameters

Bias, SE, and RMSE were computed for each parameter of each item. Table 6 shows average Bias, average SE, and average RMSE for each item parameter and for each method, where average is taken over the 240 pretest items. Table 6 presents how well the three methods recovered a particular item parameter. Performance of the methods differed for different item parameters and error indices. For simplicity, the descriptions here focus on average RMSE (total error). For the *a*-parameter, Stocking's Method B and the MEM method produced the smallest average RMSEs. The OEM method yielded a larger RMSE for the *a*-parameter. For the *b*-parameter, the MEM method produced the smallest average RMSE followed Stocking's Method B. For the *c*-parameter, the MEM method produced the smallest average RMSE and the OEM method yielded the second smallest average RMSE. Stocking's Method B produced a larger average RMSE for the *c*-parameter.

TABLE 6

Average Bias, Average Standard Error, and Average Mean Square Error

Estimation Method	Bias	*a* SE	RMSE	Bias	*b* SE	RMSE	Bias	*c* SE	RMSE
OEM	-0.1294	0.1301	0.1994	0.0367	0.1602	0.1977	0.0151	0.0289	0.0482
	(0.1432)	(0.0418)	(0.1271)	(0.1322)	(0.0911)	(0.1168)	(0.0431)	(0.0109)	(0.0265)
MEM	-0.0145	0.1655	0.1869	0.0193	0.1523	0.1798	0.0039	0.0300	0.0459
	(0.1126)	(0.0602)	(0.0945)	(0.1284)	(0.0893)	(0.1252)	(0.0413)	(0.0113)	(0.0253)
Stocking's Method B	-0.0143	0.1651	0.1853	0.0081	0.1579	0.1827	0.0144	0.0342	0.0501
	(0.1082)	(0.0587)	(0.0907)	(0.1212)	(0.0912)	(0.1208)	(0.0404)	(0.0118)	(0.0251)

( ) Standard Deviation over Pretest Items

Conclusions and Discussion

The purpose of this study was to compare and evaluate three online pretest item calibration/scaling methods (the OEM method, the MEM method, and Stocking's Method B) in terms of item parameter recovery when the pretest item response data are sparse.

The MEM method produced the smallest average total error (i.e., average WMSE) in recovering the 240 pretest item characteristic curves. The MEM method performed well in keeping the scale of the pretest items on the same scale as that of the true parameters. The MEM method also worked well in recovering individual item parameters (e.g., *a*-, *b*-, and *c*-parameters). Stocking's Method B yielded the second smallest average WMSE and resulted in pretest item parameter estimates on the scale of the true parameters. The OEM method yielded a large average WMSE. With the OEM method, the scale of the pretest item parameter estimates deviated from that of the true parameters.

Most of the results in this study were consistent with the results in Ban et al. (2000). Ban et al. (2000) reported that the MEM method performed better than Stocking Method B for all

criteria and sample sizes studied. This study showed that the MEM method also outperformed Stocking Method B, except for the estimates of the *a*-parameter.

The MEM method appears to be the good choice as a pretest item calibration method. Compared to other methods in this study, the MEM method produced the smallest parameter estimation error without requiring any anchor items. Stocking's Method B also worked well, but it requires anchor items to be seeded or larger sample sizes would be needed without anchor items (Stocking, p. 21, 1988). The OEM method, which also does not require any anchor items, produced larger error in parameter estimation than the other methods in this study.

The pretest item administration design in this study postulated that a pretest item pool exists and each examinee takes some of the pretest items where there are common pretest items administered to different blocks of examinees. The pretest items were then calibrated concurrently, although several sets of pretest items could be calibrated separately set by set. Some studies (Hanson & Béguin, 1999; Wingersky, Cook, & Eignor, 1987) reported that the concurrent item calibration produces lower parameter estimation errors than the separate calibration.

As future research, performance of the methods should be further evaluated when pretest items are poor or do not fit the 3PL model. The pretest items used in this study were operational items, so the quality of the items was high. Since the MEM method uses item responses on both the operational items and pretest items to get the posterior distribution for iterations after the first, any poor pretest item could affect the parameter estimates of other pretest items with the MEM method. Unlike the MEM method, the OEM method uses item responses only on the operational items to obtain the posterior distribution used for pretest item parameter estimation. For the OEM method, a bad pretest item would not impact the parameter estimation of the *a*'s,

b's, and *c*'s of the other items. Stocking's Method B uses the operational item responses to estimate examinees' abilities and calibrates the pretest items by fixing the ability estimates, so having a bad pretest item would only affect the parameter estimates of that particular bad pretest item. It would be worthwhile to investigate the extent to which the methods produce errors in parameter estimation when the poor pretest items exist.

References

ACT, Inc. (1997). *ACT assessment technical manual.* Iowa City, IA: Author.

Ban, J.-C., Hanson, B. H., Wang, T., Yi, Q., & Harris, D. J. (2000). *A comparative study of online pretest item calibration/scaling methods in computerized adaptive testing.* (ACT Research Report 00-11). Iowa City, IA: ACT, Inc. [Available at http://www.b-a-h.com/papers/paper0003.html].

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.

Hanson, B. A. (2000a). *Estimation toolkit for item response models (ETIRM).* [Available at http://www.b-a-h.com/software/cpp/etirm.html].

Hanson, B. A. (2000b). *Estimation program for dichotomous item response models (EPDIRM).* [Available at http://www.b-a-h.com/software/epdirm/index.html].

Hanson, B. A., & Béguin, A. A. (1999). *Separate versus current estimation of IRT item parameters in the common item equating design* (ACT Research Report 99-8). Iowa City, IA: ACT, Inc.

Haynie, K. A., & Way, W. D. (1995). *An investigation of item calibration procedures for a computerized licensure examination.* Paper presented at symposium entitled Computer Adaptive Testing, at the annual meeting of NCME, San Francisco.

Hsu, Y., Thompson, T. D., & Chen, W.-H. (1998). *CAT item calibration.* Paper presented at the Annual Meeting of the National Council on Measurement in Education, San Diego.

Mislevy, R. J., & Bock, R. J. (1990). *BILOG3: Item analysis and test scoring with binary logistic model* (2nd ed.). Mooresville, IN: Scientific Software.

Stocking, M. L. (1988). *Scale drift in on-line calibration* (ETS Research Report 88-28). Princeton, NJ: ETS.

Stocking, M. L., & Lord, F. M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7*, 201-210.

Wainer, H., & Mislevy, R. J. (1990). Item response theory, item calibration, and proficiency estimation. In Wainer, H. (Ed.), *Computer adaptive testing: A primer* (Chapter 4, pp. 65-102). Hillsdale, NJ: Lawrence Erlbaum.

Urry, V. W. (1970). *A Monte Carlo investigation of logistic test models.* Unpublished doctoral dissertation, Purdue University.

Wingersky, M. S., Cook, L. L., & Eignor, D. R. (1987). *Specifying the characteristics of linking items used for item response theory item calibration* (ETS Research Report 87-24). Princeton, NJ: ETS.